Statistical Inference in Retrieval Effectiveness Evaluation

Jacques Savoy

to appear in

Information Processing & Management

Institut interfacultaire d'informatique
Université de Neuchâtel
Pierre-à-Mazel 7
CH - 2000 Neuchâtel (Switzerland)

Abstract Evaluation methodology, and particularly its statistical tests associated, plays a central role in the information retrieval domain which maintains a strong empirical tradition. In an effort to evaluate the retrieval effectiveness of a search algorithm, this paper focuses on the average precision over a set of fixed recall values. After reviewing traditional evaluation methodology through the use of examples, this study suggests applying another statistical inference methodology called bootstrap, within which no particular assumption is needed about the distribution of the observations. Moreover, this scheme may be used to assert the accuracy of virtually any statistic, to build approximate confidence interval, and to verify whether a statistically significant difference exists between two retrieval schemes, even when dealing with a relatively small sample size. This study also suggests selecting the sample median rather than the sample mean in evaluating retrieval effectiveness where the justification for this choice is based on the nature of the information retrieval data.

Keywords: Bootstrap, goodness-of-fit test, paired t-test, stemming algorithm.

1. INTRODUCTION

Information retrieval research has a strong empirical tradition and the comparison of retrieval performances is based on test collections containing a set of documents, a set of queries and, for each request, a set of relevance judgments. Good test collections are very expensive to produce and there are some inherent characteristics, often unknown, that may favor one retrieval scheme to the detriment of another. We therefore suggest using of more than one test collection when comparing the relative performance of two retrieval schemes.

Moreover, when evaluating a retrieval scheme, we assume that the following three hypotheses are always respected (Tague-Sutcliffe & Blustein, 1992), (Hull, 1993). Firstly, all queries included in a test collection are independent or not obviously related. Secondly, all documents contained in a test collection are judged either relevant or irrelevant to a given request. Thirdly, each relevant record is equally important in satisfying the user's information need. Thus, the relevance of a given document does not depend on the number of relevant and already retrieved records.

Of course, these assumptions are not really realistic, but they can be adopted as a first approximation. For example, we recently suggested a retrieval scheme based on the relationships between past queries in order to enhance the ranking of related future requests (Savoy, 1994), and the underlying assumption of this retrieval scheme clearly contradicts the first hypothesis. The second supposition implies that a "good" test collection must include a list of pertinent documents obtained, ideally through a manual inspection of all documents contained in the corpus. However, as Fox (1983, pp. 41-42) notes:

> "... it would be a formidable task to obtain exhaustive relevance judgments. The users who provided the questions were not likely to be willing to make such an effort and even if they were, the time and expense required would have been prohibitive. ... Though exhaustive relevance information was not available, it was assumed for the sake of retrieval

> evaluation that the judgments provided were a good approximation to having complete information."

Moreover, the relevance assessments given by users are subjective, as reported by (Saracevic, 1975), (Schamber, 1994) and (Harter, 1996). Cleverdon (1984, p. 39) corroborates this finding when he writes:

> " ... if two scientists or engineers are asked to judge the relevance of a given set of documents to a given question, the area of agreement may not exceed 60 percent."

Finally, the third hypothesis is not totally realistic because a user will naturally attach a greater utility to the first retrieved and relevant document than to the 25th. However, these criticisms related to the design of test collections, and particularly to the method used to obtain relevance judgments, cannot be a well grounded argumentation to invalidate retrieval experiments (Salton, 1992).

These three hypotheses lead us to define a retrieval effectiveness measure on the basis of both the number of relevant documents and the number of retrieved records. Respecting these two criteria, the average precision at eleven standard recall values can be considered as a good retrieval effectiveness measure (Tague-Sutcliffe, 1992, pp. 483-484), (Salton, 1992), (Tague-Sutcliffe & Blustein, 1994), and this means is widely accepted throughout information retrieval literature. However, other approaches can also be considered, see (van Rijsbergen, 1979, Chapter 7), and for a broader perspective about evaluation of IR systems, see (Saracevic, 1995).

This paper is organized as follows. The first part describes typical experimental methodology used to determine whether or not a retrieval scheme is better than another. In particular, we compare two suffix-stripping algorithms and analyze the distribution of differences in the average precision at eleven recall values. The second part presents the motivations of using the sample median instead of the sample mean as a location statistic in information retrieval studies. Moreover, the bootstrap methodology is briefly introduced, and, based on this approach, we show

how to build confidence intervals and hypothesis testing by presenting simple algorithms and illustrating them by examples.

2. EVALUATION OF RETRIEVAL SYSTEM

Traditional evaluation methodology compares the average precision at eleven recall values to determine whether a search strategy is better, equal of worse than another. As examples to illustrate our purposes, this paper will evaluate two stemming algorithms.

The first section describes two suffix-stripping algorithms used throughout this study. Section 2.2 introduces two well-known retrieval models, namely the classical Boolean model and the vector-space scheme. This section will also present an informal rule to decide whether a search strategy can be judged better than another. Based on goodness-of-fit tests, Section 2.3 analyses the distribution of the difference between two observation samples. Finally, three statistical tests are applied to determine whether or not Porter's stemming scheme is better than the s-suffix stripping algorithm.

2.1 Suffix stripping algorithms

A stemming algorithm reduces inflectional and derivational variants of words to a common form. For example, the words "thinking", "thinkers" or "thinks" are reduced to the stem "think". To be precise, the root of a word is obtained by removing both suffixes and prefixes while the stem is obtained by deleting only the suffixes.

In information retrieval, grouping words having the same root under the same stem (or indexing term) will increase the success of matching of documents to a query (van Rijsbergen, 1979, Chapter 2). Therefore, such an automatic procedure may be a valuable tool for the enhancement of the retrieval effectiveness, under the assumption that words having the same stem refer to the same idea or concept.

To define such a procedure, a minimal stemming procedure called s-suffix will only conflate the singular and plural word forms according to the following three rules:

1. if a word ends in «-ies», but not «-eies» or «-aies»
   then replace «-ies» by «-y»;
2. if a word ends in «-es», but not «-aes», «-ees» or «-oes»
   then replace «-es» by «-e»;
3. if a word ends in «-s», but not «-us» or «-ss»
   then remove «-s».

More sophisticated stemming schemes have already been proposed for suffix removal (Lovins, 1968), (Porter, 1980), (Savoy, 1993), (Krovetz, 1993), (Paice, 1994), (Hull, 1996). Even if the retrieval effectiveness of such procedures has already been analyzed (Frakes, 1992), the current study is concerned with a statistical analysis of the difference between two retrieval schemes, and the evaluation of the s-suffix and Porter's suffix-stripping algorithms will be used to illustrate our purposes.

2.2 *Experimental results*

As a retrieval effectiveness measure for a given query i, we suggest incorporating the average precision at eleven standard recall values, measure noted x_i. For a sample of n queries, we may compute the average precision and its estimated standard error $\hat{\sigma}_{\bar{x}}$ according to the following formulae:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{S^2}{n}} = \frac{S}{\sqrt{n}} \qquad \text{with } S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (1)$$

During our evaluations, we retain 50 requests for the CACM collection, because two queries (#2 and #57) are addressed to external descriptors of documents (e.g., author's name). The CISI corpus contains 35 requests, both in natural language and Boolean forms.

As a first retrieval strategy, we adopt the classical Boolean scheme within which the retrieved records are ranked according to the decreasing order of their publication date. This ranking scheme favors recent documents and simulates a presentation order common in commercial systems.

As a second retrieval scheme, we will consider the vector-space model, within which the retrieval status value (RSV) of each retrieved record is computed according to the well-known cosine similarity measure (Salton, 1989) as follows:

$$\mathrm{RSV}(D_i, Q) = \frac{\sum_{j=1}^{q} w_{ij} \cdot w_{qj}}{\sqrt{\sum_{j=1}^{t} w_{ij}^2 \cdot \sum_{j=1}^{q} w_{qj}^2}}$$

where w_{ij} represents the weight of term T_j in document D_i, w_{qj} the weight of term T_j in the current request Q, t the number of indexing terms in the collection, and q the number of stems contained in the query Q.

To assign a value to both w_{ij} and w_{qj}, our automatic indexing procedure applies the following formula:

$$w_{ij} = ntf_{ij} \cdot nidf_j \qquad \text{where } ntf_{ij} = \frac{tf_{ij}}{\max tf_i} \qquad \text{and } nidf_j = \log\left(\frac{m}{df_j}\right) / \log(m)$$

where tf_{ij} is the frequency of the term T_j in the document D_i, m the number of documents D_i in the collection, df_j the number of documents in which T_j occurs, and idf_j the inverse document frequency.

The evaluation of the Boolean and vector-space models according to our two stemming procedures is shown in Table 1.

	Precision (% change)	
Model \ Collection	CACM	CISI
Classical Boolean model s-suffix	19.52	13.08
Classical Boolean model Porter's algorithm	21.00 (+7.6%)	13.66 (+4.4%)
Vector-space model (cosine) s-suffix	30.92	18.60
Vector-space model (cosine) Porter's algorithm	32.58 (+5.4%)	20.28 (+9.0%)

Table 1: Average precision at eleven standard recall values

To decide whether a search strategy is better than another, we need a decision rule. To define such a rule, we may use the following rule of thumb: a difference of at least 5% in average precision is generally considered significant and a 10% difference is considered very significant (Sparck Jones & Bates, 1977, p. A25).

According to this rule, and for the CACM collection, we may conclude that Porter's stemming algorithm performs with significant enhancement for both retrieval schemes. When inspecting results obtained with the CISI corpus, the vector-space model shows a significant increase when using Porter's procedure. The Boolean model however depicts only a near significant enhancement. Based on these results, it seems more appropriate to choose Porter's stemming algorithm rather than the s-suffix scheme.

However, we may wish to establish that the difference in retrieval effectiveness under two conditions is statistically different or that the difference does not simply occur by chance. To achieve this goal, we may base our decision rule on either parametric or nonparametric tests instead of applying the described informal rule.

2.3 *Goodness-of-Fit tests*

The most important source of information in defining whether a search scheme is more effective than another, is the study of the difference in average precision at eleven recall values. Such a measure is computed as follows:

$$x_i^d = x_i^a - x_i^b \tag{2}$$

in which x_i^{a} and x_i^{b} represent the average precision for the *i*th request obtained with the retrieval strategy a, respectively strategy b. In this study, the condition a indicates the average precision using the s-suffix algorithm while condition b reflects the result obtained with Porter's stemming procedure.

Since parametric tests generally assume that the random variable x_i^{d} follows a Gaussian distribution, we want to verify such an assumption using the well-known chi-square test (Freeman, 1987), (Conover, 1980, Section 4.5) instead of following our prior feelings.

In such tests, the null hypothesis H_0 states that the underlying distribution of the observed data follows a given probability law, the normal distribution in our case. Under this condition, we compute a statistic based on the hypothesized distribution and the probability of observing this value. If the resulting probability is less than a specified significance level α, we may conclude that the empirical distribution of the data and the hypothesized distribution are statistically different, or that the difference between both distributions could not only have occurred by chance. In the given context, the following two hypotheses are:

H_0: "The distribution function of the observations is a Gaussian distribution function"

H1: "The distribution function of the observed variable is different from the normal"

After grouping the observations into c classes, the Pearson chi-square statistic X^2 will compare the observed count O_k and the expected number of observations E_k in each cells k = 1, 2, ..., c (Conover, 1980, Section 4.5), (Freeman, 1987). The expected number E_k is computed according to the hypothesized distribution provided by H_0. Moreover, we generally require that the expected count in any class to be at least 1.0, and more than 5.0 in at least 80 percent of the cells. To respect these criteria, 10 groups were defined for the CACM collection, and 7 classes for the CISI corpus. The Pearson chi-square statistic is computed according to the following formula:

$$X^2 = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i}$$

and the value X^2 follows a chi-square distribution with c - p - 1 degrees of freedom where c means the number of classes, and p the number of estimated parameters, two for a Gaussian distribution.

For the CACM collection, the number of degrees of freedom is 10 - 2 - 1 = 7 leading to the critical values $t_\alpha = 14.067$ with a significance level $\alpha = 0.05$, or $t_\alpha = 18.475$ with $\alpha = 0.01$ (the notation t_α does not mean that the underlying distribution is a Student). For the CISI corpus, the number of degrees of freedom is 7 - 2 - 1 = 4 and the critical values t_α are 9.488 with a significance level $\alpha = 0.05$, or $t_\alpha = 13.28$ with $\alpha = 0.01$. The resulting decision rule consists of rejecting the null hypothesis if the value of X^2 exceeds t_α (one-sided test).

Besides this test, we may also consider the log likelihood statistic G^2, or the weighted least squares Q. The computation of these two tests are based on the following equations, and both statistics follow a chi-square distribution with c-p-1 degrees of freedom.

$$G^2 = 2 \cdot \sum_{i=1}^{c} O_i \cdot \ln\left(\frac{O_i}{E_i}\right) \qquad \text{and} \quad Q = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{O_i}$$

For the latest statistic, when the value O_i becomes zero for a given cell, we replace it by 0.5. The values of these three statistics are reported in Table 2 and are based on the difference between the s-suffix algorithm and Porter's stemming procedure.

Statistics	CACM Boolean	CACM vector	CISI Boolean	CISI vector
Pearson chi-square	$X^2 = 152.4$ $t_{\square} = 18.475$ H_0 rejected $\square = 0.01$	$X^2 = 28.8$ $t_{\square} = 18.475$ H_0 rejected $\square = 0.01$	$X^2 = 25.6$ $t_{\square} = 13.28$ H_0 rejected $\square = 0.01$	$X^2 = 10.8$ $t_{\square} = 9.488$ H_0 rejected $\square = 0.05$
Likelihood	$Q = 88.306$ $t_{\square} = 18.475$ H_0 rejected $\square = 0.01$	$Q = 47.898$ $t_{\square} = 18.475$ H_0 rejected $\square = 0.01$	$Q = 36.856$ $t_{\square} = 13.28$ H_0 rejected $\square = 0.01$	$Q = 21.905$ $t_{\square} = 13.28$ H_0 rejected $\square = 0.01$
Weighted least squares	$G^2 = 85.234$ $t_{\square} = 18.475$ H_0 rejected $\square = 0.01$	$G^2 = 27.859$ $t_{\square} = 18.475$ H_0 rejected $\square = 0.01$	$G^2 = 24.143$ $t_{\square} = 13.28$ H_0 rejected $\square = 0.01$	$G^2 = 11.413$ $t_{\square} = 9.488$ H_0 rejected $\square = 0.05$

Table 2: Goodness-of-Fit tests

From these results, we may conclude that, for the CACM collection, the distribution of the difference for both the Boolean scheme and the vector-space model does not follow a Gaussian distribution at a significance level of 0.01. For the CISI corpus, we must reject the null hypothesis H_0 for the classical Boolean model (significance level of 0.01). For the vector-space model, the Pearson chi-square and the weighted least squares tests lead to the conclusion that the hypothesis H_0 must be rejected at the significance level of 0.05, but cannot be rejected for $\square = 0.01$. However, not all retrieval performance differences follow this pattern. For example, when comparing the Boolean model and the vector-space scheme, we cannot reject the hypothesis that the distribution of the difference between these two schemes follows a normal distribution.

Besides the chi-square tests, we may also use the more powerful Kolmogorov-Smirnov method. The basic idea of this statistic test consists of computing the absolute value of the difference between $F_n(x)$, the empirical distribution of the observations, and $F(x)$ the theoretical distribution provided by the null hypothesis H_0 (Conover, 1980, pp. 346-353).

To verify whether the difference values comply with a Gaussian distribution, we formulate the null hypothesis H_0 that the underlying distribution is normal and we calculate the following statistic:

$$T = \max_{x} \; [\mid F_n(x) - F(x) \mid]$$

in which T measures the absolute value of the maximum deviation between F_n and F. The results depicted in Table 3 are based on a two-sided test, and the null hypothesis H_0 is rejected if the value of the statistic T is greater than the value specified by t□ (since the underlying distribution is symmetrical, only one percentile value is depicted in Table 3) .

Statistic	CACM Boolean	CACM vector	CISI Boolean	CISI vector

Kolmogorov	T = 0.357168
	t□ = 0.2305
	n = 50
	H_0 rejected
	□ = 0.01
		T = 0.19381
		t□ = 0.1923
		n = 50
		H_0 rejected
		□ = 0.05
			T = 0.277072
			t□ = 0.269
			n = 35
			H_0 rejected
			□ = 0.01
				T = 0.172868
				t□ = 0.224
				n = 35
				H_0 cannot be rejected: □=0.05

Table 3: Summary of Kolmogorov-Smirnov test

The conclusions that can be drawn from Table 3 confirm those of the chi-square tests. For the CACM corpus, the difference in average precision does not follow a Gaussian distribution. Except for the vector-space model and using the CISI collection, the Kolmogorov-Smirnov test indicates that we cannot reject the fact that the distributions of the data may follow a normal distribution (significance level of 0.05). Since the Gaussian distribution is an assumption for various parametric tests, and following (van Rijsbergen, 1979), we conclude that these tests are suspect in most information retrieval contexts.

2.4 *Statistical tests*

The aim of statistical tests is to know whether or not the difference between two retrieval schemes is really significant or if this difference could have occurred by chance. The resulting decision is strengthened (1) when the difference values are relatively high; or (2) when these values are, more or less, always in favor of the same retrieval scheme; and (3) when the sample size grows.

Based on the results of Table 1, we have informal evidence that Porter's stemming scheme is better than the s-suffix approach. Using statistical tests, we want to confirm this conclusion. Therefore, in the following tests, the null hypothesis

H_0 states that the s-suffix scheme is better or equal to Porter's stemming algorithm. Such a null hypothesis plays the role of a devil's advocate, and we hope that the resulting statistic will lead us to reject this hypothesis (one-sided test).

Under H_0, each test computes a statistic T and calculates the achieved significance level of this test which is the probability of observing a value at least as extreme as T when the null hypothesis H_0 is true. If this probability is less than a specified significance level α, we may conclude that the search schemes are significantly different.

In the following computations, the results corresponding to the retrieval scheme a are obtained with the s-suffix algorithm, while Porter's stemming procedure represents condition b.

The "paired t test" represents the first statistical test that we might consider, under the assumption that the difference x_i^d follows a normal distribution. The formulation of the statistic T is described by Equation 3 (Conover, 1980, pp. 290-292). The corresponding decision rule consists of rejecting H_0 if $T < t_\alpha$, where the percentile t_α follows a Student's distribution with n-1 degrees of freedom.

$$T = \frac{\bar{x}^d}{S_{x^d} / \sqrt{n}} \qquad \text{with } \bar{x}^d = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^d \quad \text{and } S_{x^d} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i^d - \bar{x}^d\right)^2} \qquad (3)$$

The result of this parametric test is of questionable value, because, as we have shown in Section 2.3, the distribution of the difference does not follow a Gaussian distribution, and this is a problem only for small sample which is the case in our study. As mentioned in (van Rijsbergen, 1979), this test seems therefore inappropriate.

However, even if the distribution of the observations is not normally shaped but if the empirical distribution is roughly symmetric, the t-test can be still a useful test because it is relatively robust in the sense that the indicated significance level is not far from the true α level. However, testing symmetry is a more complex procedure, e.g., (Antille *et al.*, 1982).

The Wilcoxon Signed-Ranks test is based on the statistic T computed according to Equation 4 (Siegel, 1956, pp. 75-83), (Conover, 1980, pp. 280-288). If the null hypothesis H_0 is true (the s-suffix scheme exhibits a better or equal effectiveness than Porter's stemming procedure), the values of T tend to be large, and small values of T indicate that H_0 is false. Therefore, our decision rule is to reject H_0 if $T < z_\alpha$, where the percentile z_α follows a standard normal distribution N(0,1). However, the approximation, $T \overset{\cdot}{\sim} N(0,1)$, is valid only if $n > 20$, which is the case in the current study. Finally, for both the Wilcoxon and Sign tests, ties (if any) are removed before the underlying statistics are computed, and the resulting size n is indicated in Table 4.

$$T = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}} \qquad \text{with } R_i = \operatorname{sign}(x_i^{\,d}) \cdot \operatorname{rank} \mid x_i^{\,d} \mid, \text{ and } T \overset{\cdot}{\sim} N(0,1) \qquad (4)$$

For the Sign test (Siegel, 1956, pp. 68-7)], (Conover, 1980, pp. 122-129), the corresponding statistic is expressed in Formula 5 and represents the number of times that the s-suffix stemming algorithm returns a better performance than Porter's stemming scheme.

$$T = \sum_{i=1}^{n} I[\bar{x}^{d} > 0] \qquad \text{with } I[\bar{x}^{d} > 0] = \begin{cases} 1 & \text{if } \bar{x}^{d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } t_\alpha = \frac{1}{2} \cdot (n + z_\alpha \cdot \sqrt{n}\ ) \quad \text{with } z \sim N(0,1), \text{ and } n > 20 \qquad (5)$$

Large values of T indicate that the average precision provided by the s-suffix procedure tends to be higher than Porter's stemming algorithm, as assumed by H_0. The decision rule will compare the values of T and t_α, and the null hypothesis H_0 will be rejected if $T \leq t_\alpha$.

The results of Table 4 do not corroborate the conclusions based on our rule of thumb described in Section 2.2. For all statistics, except two, the decision is to accept the null hypothesis H_0 stating that the s-suffix scheme presents better, or at least equal, retrieval effectiveness compared to Porter's stemming procedure. However, a decision to accept H_0 is not equivalent to the opinion that the null hypothesis H_0 is true, but, instead, represents the fact that "H_0 has not been shown to be false" resulting of insufficient evidence against H_0.

As mentioned, both Wilcoxon and Sign test are based on a reduced sample size because ties are removed. A closer look at the Table 4 data demonstrates that, for the CACM collection and using the classical Boolean model, these tests are based only on 24 pairs of values, leading to the conclusion that, for 26 requests, both suffix-stripping schemes return identical retrieval performances.

The previous tests are based on average precision differences. However, if the difference is relatively small for a given request (e.g., less than 0.1%, an arbitrary level), we might consider such a value to be zero as suggested by the following equation. For example, if under two different conditions and for one query, we obtain 25.33% and 25.27% as average precision, it seems reasonably to view this difference as non significant. Ignoring these very small differences, the computation of the various statistical tests leads to the same conclusion.

$$x_i{}^d = \begin{cases} x_i^a - x_i^b & | x_i^a - x_i^b | > 0.1 \\ 0 & \text{otherwise} \end{cases}$$

When comparing the conclusions drawn for the rule of thumb given in Section 2.2 and those from the statistical tests, we are faced with different deductions. For example, for the CISI collection, the average precision of the vector-space model based on Porter's stemming scheme is 9% higher than those provided by the s-suffix algorithm (20.28 vs. 18.6 in Table 1). According to the informal rule, Porter's algorithm results in a significant enhancement over the s-suffix scheme. However, such an affirmation is not confirmed by all statistical tests.

Statistics	CACM Boolean	CACM vector	CISI Boolean	CISI vector

Paired t-test	T = -1.4301
	t_α = -1.6775
	n = 50
	H_0 cannot be rejected: α=0.05
		T = -0.8107
		t_α = -1.6775
		n = 50
		H_0 cannot be rejected: α=0.05
			T = -0.8355
			t_α = -1.697
			n = 35
			H_0 cannot be rejected: α=0.05
				T = -2.2044
				t_α = -1.697
				n = 35
				H_0 rejected
				α = 0.05

Wilcoxon	T = -1.0
	z_α = -1.6449
	n = 24
	H_0 cannot be rejected: α=0.05
		T = -1.877
		z_α = -1.6449
		n = 48
		H_0 rejected
		α = 0.05
			T = 0.3723
			z_α = -1.6449
			n = 31
			H_0 cannot be rejected: α=0.05
				T = -1.4905
				z_α = -1.6449
				n = 35
				H_0 cannot be rejected: α=0.05

Sign test	T = 12
	t_α = 7.97
	n = 24
	H_0 cannot be rejected: α=0.05
		T = 19
		t_α = 18.30
		n = 48
		H_0 cannot be rejected: α=0.05
			T = 19
			t_α = 10.92
			n = 31
			H_0 cannot be rejected: α=0.05
				T = 16
				t_α = 12.63
				n = 35
				H_0 cannot be rejected: α=0.05

Table 4: Results of statistical tests

To clarify this dilemma, we suggest the following explanation. The pragmatic rule is based on the sample mean which, by definition, for various requests, may hide the fact that the s-suffix scheme may return better results than Porter's stemming algorithm. The statistic T of the Sign test in Table 4 indicates that for 16 queries over 35, the s-suffix returns a better retrieval performance. However, when Porter's stemming scheme gives a better performance, the difference is relatively high, and leads to an average precision greater than the mean provided by the s-suffix procedure. Thus, an informal rule, based on an overall measure, leads to the conclusion that Porter's stemming scheme is better, while the Sign test, grounded on all individual requests, does not corroborate to this finding.

According to van Rijsbergen (1979), we know that the conditions required for the application of these tests are not really met in the information retrieval context.

> "[The Wilcoxon Matched-Pairs] test is done on the differences $D_i = Z_a(Q_i) - Z_b(Q_i)$, but it is assumed that D_i is continuous and that it is derived from a *symmetric* distribution, neither of which is normally met in IR data. ... [The sign test] makes no assumptions about the form of the underlying distribution. It does, however, assume that the data are derived from a continuous variable and that the $Z(Q_i)$ are *statistically independent*. These

> two conditions are unlikely to be met in a retrieval experiment. Nevertheless given that some of the conditions are not met it can be used *conservatively*." (van Rijsbergen, 1979, pp. 178-179)

However, for (Hull, 1993), even if these conditions are not strictly respected, statistical tests are still valid, because with a sufficiently large sample, discrete and bounded measures are often well-approximated by continuous distribution. Our purpose is not to elaborate on these two positions, but rather to go beyond this debate in suggesting another statistical methodology which relieves the investigator of having to make assumptions underlying both parametric and nonparametric statistical models (e.g., variances for different treatment groups are approximately equal, effects and interactions of the independent variables are additive, etc.).

3. ANOTHER LOOK AT RETRIEVAL EFFECTIVENESS MEASURES

The distribution of the difference between two retrieval schemes does not follow a normal distribution in all circumstances, and, therefore, parametric tests are of doubtful value. Nonparametric tests stipulate hypotheses that may not hold in the context of information retrieval analysis (van Rijsbergen, 1979), and other statistical models may be based on unrealistic or unverifiable assumptions. Moreover, the average precision does not represent the only measure available to quantify the difference between two retrieval schemes, and we may also consider the median, a more robust location statistic. Based on the bootstrap paradigm, this chapter suggests another approach for deciding whether or not a retrieval strategy provides a better retrieval effectiveness than another.

The rest of this paper is organized as follows. Section 3.1 analyzes the choice of the sample mean as a summary statistic for information retrieval evaluation studies and suggests replacing this measure by the sample median. After obtaining a point estimator, we need some indication of its accuracy or a numerical value of its standard error. To achieve this goal, Section 3.2 explains the basic idea of the bootstrap approach and illustrates its use through examples. Section 3.3 describes how one can build approximate confidence intervals based on the bootstrap paradigm, while the last section describes the algorithm that can be apply for inferential decisions.

3.1 *Choice of a summary statistic*

Sample mean is one approach measuring the central tendency of a distribution, and the sample median approach can also be considered. This latter location measure is represented by the value in the middle of the sorted sample. For example, the median and the mean $\bar{x}$ of the sample X = {25, 26, 27, 28, 29} are both equal to 27. When a sample contains an even number of values, the median can be computed according to the average between the two values in the middle of the sample. However, what are the differences between these two statistics?

When we introduce a zero value in our sample leading to X = {0, 25, 26, 27, 28, 29}, the sample median is evaluated as (26 + 27) / 2 = 26.5. The introduction of this value zero, represents a very different value compared to the rest of the sample, leading to a sample mean value of 22.5. Such a number does not seem typical; after all, 5 of the 6 values are greater than 22.5. Since the sample mean is sensitive to the presence of extreme scores, it is not a particularly good measure of location when the distribution is skewed and / or truncated. By contrast, the median is not dramatically changed by the new value, and therefore represents a more robust summary statistic.

In the information retrieval domain, the presence of an average precision of 0 for a request reflects the fact that the retrieval scheme is unable to find any relevant record. Particularly when an retrieval scheme is based on the Boolean model, such a phenomena can be explained by various reasons as, for example, by writing a too restrictive query, introducing a spelling error or a variant of a term (e.g., "foetus", "fetus"), confusion between the operator AND and OR (e.g., the request "whale" OR "dolphin" was written as "whale" AND "dolphin").

In a less extreme situation, the analysis of various TREC experiments demonstrates that a retrieval scheme may perform very well for some queries and poorly for other requests. In Section 2.4., we also find that the overall statistics, like the average precision, may hide performance irregularities among requests when comparing two retrieval schemes. Moreover, we know that the vector-space model does not perform very well with a very short query. From these considerations, we really need a robust summary statistic to evaluate the retrieval effectiveness and the median seems to be a better measure than the mean.

However, the sample mean displays interesting properties. We know of both its expectation and variance which provides a general idea of the accuracy of this point estimator. Moreover, based on the central limit theorem, the distribution of $\bar{x}$ will be approximately normal as the sample size n gets larger. Thus, we may write:

$$\bar{x} \stackrel{\cdot}{\sim} N\left(\mu ; \frac{\hat{\sigma}}{\sqrt{n}}\right) \qquad (6)$$

in which the symbol $\stackrel{\cdot}{\sim}$ means that the random variable follows approximately a Gaussian distribution. This formulation leads us to construct confidence intervals around the observed sample mean and hypothesis testing. However, such an approximation can be good if n is large, but can be quite inaccurate for the sample size actually available. Moreover, other location statistics are not necessary represented by a neat formula like Equation 6, and the assessment of the accuracy of such an estimator can be quite hard. Therefore, some authors (e.g., (Grimm, 1993)) suggest that the median must be considered only as a descriptive measure.

Collection and Model	n	$\bar{x}$	median	S
CACM Boolean model, s-suffix	50	19.52	12.01	22.62
CACM Boolean model, Porter	50	21.00	14.51	22.63
CISI Boolean model, s-suffix	35	13.08	10.92	9.79
CISI Boolean model, Porter	35	13.66	13.53	9.65
CACM vector-space model, s-suffix	50	30.92	26.99	20.78
CACM vector-space model, Porter	50	32.58	29.86	21.46
CISI vector-space model, s-suffix	35	18.60	17.31	12.58
CISI vector-space model, Porter	35	20.28	20.88	13.10

Table 5: Average precision statistics at eleven standard recall values

In order to obtain a picture of the value of the median for our retrieval schemes, Table 5 depicts the statistics for both test collections. Limited to a point estimate, a given statistic is not especially interesting, and we really need some indication about the accuracy of such a value. To achieve this objective, the bootstrap method can be a useful tool.

3.2 Principles and examples of bootstrap methodology

To assign measures of accuracy to virtually any statistical estimators, we suggest using the bootstrap methodology (Efron & Tibshirani, 1986), (Léger *et al.*, 1992), (Efron & Tibshirani, 1993). Within this paradigm, we do not have to rely entirely on the central limit theorem to obtain a numerical value of estimator accuracy. Such an approach is very attractive in the information retrieval domain, because we could use the median instead of the mean to measure the central tendency of a sample.

The basic idea of the bootstrap approach is simple and can be explained as follows. For retrieval effectiveness measures, we have a sample of observations X = $\{x_1, x_2, ..., x_k, ..., x_n\}$ of size n, drawn from a population of queries possessing a probability distribution F. If we know the real distribution F, we may compute the underlying parameter of interest, e.g., the median or the mean, according to □ = t(F). Since the distribution F is unknown, we want to estimate the parameter □ by a point estimate $\hat{\square}$ = t($\hat{F}$) based, for example, on the plug-in principle. Within this approach, the estimate is computed according to the same function, t() in our case, which should be applied if we know the real distribution F. In this computation, we substitute F by the empirical distribution $\hat{F}$. Traditional statistics theories may provide other functions for obtaining a point estimate based on a sample of observations. However, the aim of the bootstrap methodology is not to provide another formula to calculate an estimator, but to achieve a measure of accuracy of any statistical estimate.

In order to achieve this goal, the computer generates a set of bootstrap samples $X^{*i} = \{x^*_1, x^*_2, ..., x^*_k, ..., x^*_n\}$, for i = 1, 2, ... B, by random sampling with replacement from X. This process garantees that each value x^*_k is mutually independent of each other and identically distributed (i.i.d.) of $\hat{F}$, where $\hat{F}$ represents the empirical distribution function putting probability 1/n on each value x_k (nonparametric bootstrap). Each bootstrap sample X^{*i} contains members of the sample X, some appearing zero times, some once, some twice, etc.

When we randomly generate a bootstrap sample, each value x_k has the probability 1/n to be selected, according to the empirical distribution $\hat{F}$. Thus, the probability that a given value x_k does not appear in a bootstrap sample of size n is:

$$p_n = \left(1 - \frac{1}{n}\right)^n, \quad \text{and when } n \to \infty, \text{ the value } p_n \to e^{-1} \approx 0.368$$

On the other hand, the probability that a given value x_k appears in a bootstrap sample is:

$$1 - p_n = 1 - \left(1 - \frac{1}{n}\right)^n, \quad \text{and when } n \to \infty, \text{ the value } 1 - p_n \to 1 - e^{-1} \approx 0.632$$

From each sample X^{*i}, we may compute $\hat{\theta}^{*i}$, the bootstrap replication of $\hat{\theta}$ computed according to the same function which was applied to compute $\hat{\theta}$ from X. Finally, the value of $\hat{\sigma}^{*}(\hat{\theta}^{*})$, the estimator of standard error associated with $\hat{\theta}^{*}$, can be considered as a good approximation to numerical value of the standard error associated with $\hat{\theta}$ (see underlying formulae in Figure 1). Of course, one bootstrap sample is clearly not enough to obtain an accurate estimate of the standard error, and the computer will repeat this procedure B times according to the algorithm depicted in Figure 1.

However, the bootstrap method may sometimes fail to give an appropriate numerical value of the standard error of an estimator θ, as mentioned by Efron & Tibshirani (1993, p. 81):

> "The difficulty occurs because the empirical distribution $\hat{F}$ is not a good estimate of the true distribution F in the extreme tail. ... The nonparametric bootstrap can fail in other examples in which θ depends on the smoothness of F."

Such cases have not been encountered in our case, within which the point estimator is a summary statistic (e.g., the sample mean or the sample median, as demonstrated in (Bickel & Freedman, 1981)). For example, it could be misleading to apply bootstrap methodology, at least as described in this paper, to estimate the standard error associated to the maximum of a distribution (Bickel & Freedman, 1981, Section 5).

```
    Estimator accuracy measurement for $\hat{\theta}$
    Given $X = \{x_1, x_2, ..., x_k, ..., x_n\}$ , a sample of size n;
    for i = 1, 2, ..., B (e.g., B = 50 to 1000)
        Generate $X^{*i} = \{x^*_1, x^*_2, ..., x^*_k, ..., x^*_n\}$ a bootstrap sample drawn
            with replacement from X, and with $x^*_k$ ~ i.i.d. of $\hat{F}$;
        Compute the statistic $\hat{\theta}^{*i}$ corresponding to each bootstrap
            sample $X^{*i}$;
    next i
    Compute $\hat{\sigma}^*(\hat{\theta}^*) = \sqrt{\frac{1}{B-1} \cdot \sum_{k=1}^{B} (\hat{\theta}^{*k} - \bar{\theta}^*)^2}$    with $\bar{\theta}^* = \frac{1}{B} \cdot \sum_{k=1}^{B} \hat{\theta}^{*k}$
```

Figure 1: General bootstrap algorithm

As an example illustrating our purposes, we will evaluate the accuracy of both the sample mean and the sample median of a retrieval scheme operating under two different conditions. Given a set of seven queries, we obtain the following results under circumstances a: {98, 70, 49, 47, 19, 11, 8}, and the following values when using method b: {73, 52, 36, 25, 20, 15, 5}. We assume that each x_k ~ i.i.d. of the distribution F, which means that the requests are not related. The standard errors for both the sample mean and median are depicted in Table 6.

Statistics \ B =	50	200	500	2000	5000	
$\hat{\sigma}_{\text{mean a}}$	12.7403	11.2162	11.4564	11.7848	11.5548	11.633
$\hat{\sigma}_{\text{median a}}$	17.9733	18.1793	18.4362	18.7467	18.904	18.841
$\hat{\sigma}_{\text{mean b}}$	8.9398	8.0931	8.0571	8.2796	8.1905	8.216
$\hat{\sigma}_{\text{median b}}$	10.2776	10.5954	11.1351	11.3062	11.5838	11.868

Table 6: Standard errors of the sample mean and the sample median

In the last column of Table 6, we have included the exact value of both standard errors. For the median, these values are obtained according to the argument that follows. For a given bootstrap sample of size seven, the probability that the median will be equal to x_i (or the fourth member of the sorted sample) is the following (Efron & Tibshirani, 1993, p.16):

$$p(i) = \sum_{j=0}^{3} \left\{ Bi(j; n; (i-1)/n) - Bi(j; n; i/n) \right\} \quad \text{with } Bi(j; n; p) = \binom{n}{j} \cdot p^{j} \cdot (1-p)^{n-j}$$

in which Bi(j; n; p) represents the binomial probability.

The resulting values, p(1)=0.0102, p(2)=0.0981, p(3)=0.2386, p(4)=0.3062, p(5)=0.2386, p(6)=0.0981, p(7)=0.0102, are used to obtain the corresponding values of the median and its associated standard error in Table 6.

From this table, one can see that as the constant B gets larger, the accuracy of the standard error gets closer to the limit value (last column). However, the particular value can sometimes be larger, sometimes smaller than this limit (e.g., see $\hat{\sigma}_{\text{mean } a}$). In another case, the values increase slightly as B grows (e.g., see $\hat{\sigma}_{\text{median } a}$) but the sampling variability decreases as B increases, or in other words, the estimates came closer and closer to the limit value. Moreover, for both samples, the standard error associated with the sample median is larger than the numerical value of the standard error of the mean. For these samples, the mean is therefore more accurate than the median.

Such a computer application however must be based on a "good" uniform pseudo-random number generator. For this purpose, we have implemented a generator grounded on combined Tausworthe sequences producing a period length about 10^{18}. Moreover, a battery of 21 statistical tests applied to this generator does not reveal any regularity in the resulting sequences (Tezuka & L'Écuyer, 1991), (L'Écuyer & Côté, 1991), even if a truly random sequence cannot be computed (Herring & Palmore, 1995).

3.3 *Approximate bootstrap confidence interval*

After obtaining a numerical value of the standard error for a given statistic, we want to go a step further in defining an approximate confidence interval for a point estimate. If we already know the value of the statistic $\hat{\theta}$ and its standard error, we actually do not know the distribution followed by $\hat{\theta}$. If the sample mean is chosen as a summary statistic, and if n is relatively large, then the following statistic Z becomes a standard normal distribution.

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \;\dot{\sim}\; N(0,1) \tag{7}$$

and, according to Equation 1, we may estimate the standard error of the sample mean as $\hat{\sigma}_{\bar{x}} = \hat{\sigma}/\sqrt{n}$. Using the standard normal distribution to find the numerical value of percentiles z_{α} and $z_{1-\alpha}$, the following standard interval estimate will have a coverage probability equal to $1 - (2 \cdot \alpha)$.

$$\mathrm{Prob}_F \left\{ z_{\alpha} \le \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \le z_{1-\alpha} \right\} = 1 - 2 \cdot \alpha$$

which can be written as:

$$\mathrm{Prob}_F \left\{ -z_{1-\alpha} \le \frac{\mu - \bar{x}}{\sigma_{\bar{x}}} \le -z_{\alpha} \right\} = \mathrm{Prob}_F \left\{ \bar{x} - z_{1-\alpha} \cdot \sigma_{\bar{x}} \le \mu \le \bar{x} - z_{\alpha} \cdot \sigma_{\bar{x}} \right\}$$

and we obtain the following confidence interval:

$$[\, \bar{x} - z_{1-\alpha} \cdot \hat{\sigma}_{\bar{x}} \,;\; \bar{x} - z_{\alpha} \cdot \hat{\sigma}_{\bar{x}} \,] \tag{8}$$

Improvement of such interval estimates can be obtained for a moderate sample size n. The assumption leading to Equation 7 is valid only if n is relatively large, and Gosset (1908) derives a better estimate when $\hat{\theta} = \bar{x}$ and demonstrates that the underlying statistic Z follows a Student's distribution with n-1 degrees of freedom. Thus, Formula 8 can be written as:

$$[\, \bar{x} - t_{n-1,1-\alpha} \cdot \hat{\sigma}_{\bar{x}} \,;\; \bar{x} - t_{n-1,\alpha} \cdot \hat{\sigma}_{\bar{x}} \,] \tag{9}$$

in which $t_{n-1,1-□}$ or $t_{n-1,□}$ follows a Student's distribution with n-1 degrees of freedom.

To obtain an interval estimate for the median and without relying entirely on the central limit theorem, we may use the following procedure. Given $X = \{x_1, x_2, ..., x_n\}$ a sample of n values representing the average precision at eleven standard recall value, we assume that each x_i ~ i.i.d. of F. To obtain an interval estimate or a confidence interval for the estimator $\hat{□}$, we will use a nested bootstrap procedure as shown in Figure 2.

In a first step, we compute the estimate $\hat{□}$ based on the available sample X. In order to obtain a numerical value of the standard error of $\hat{□}$, we generate B_1 bootstrap samples drawn with replacement from X, and such that each x^*_k ~ i.i.d. of $\hat{F}$. Based on a given bootstrap sample X^{*i}, we may compute the bootstrap replication of $\hat{□}$ noted $\hat{□}^{*i}$. To obtain a numerical value of the estimate of the standard error associated with each $\hat{□}^{*i}$, we apply a second time the bootstrap algorithm, generating B_2 bootstrap samples X^{**j} which members are drawn with replacement from X^{*i}. After B_2 draws, we may compute the numerical value of the standard error of $\hat{□}^{*i}$, noted $\hat{□}^{**}(\hat{□}^{*i})$ which is used to calculate t^*_i, a Studentized value, as:

$$t^*_i = \frac{\hat{□}^{*i} - \hat{□}}{\hat{□}^{**}(\hat{□}^{*i})}$$

which forms a sample of B_1 values used to build our estimate of the distribution of the percentile $t_□$ directly from the observations.

```
Confidence interval around $\hat{\square}$
Compute $\hat{\square} = t(\hat{F})$, the estimator based on the sample X;
for i = 1, 2, ..., $B_1$  (e.g., $B_1$ = 200 to 1000)
        Generate $X^{*i} = \{x^*_1, x^*_2, ..., x^*_k, ..., x^*_n\}$;
        Compute the statistic $\hat{\square}^{*i}$ based on the bootstrap sample $X^{*i}$;
        for j = 1, 2, ..., $B_2$  (e.g., $B_2$ = 25 to 200)
                Generate $X^{**j} = \{x^{**}_1, x^{**}_2, ..., x^{**}_k, ..., x^{**}_n\}$ a bootstrap
sample
                        drawn from $X^{*i}$ and with $x^{**}_k \sim$ i.i.d. of $\hat{F}^*$;
                Compute $\hat{\square}^{**j}$ the statistic of the bootstrap sample $X^{**j}$;
        next j
        Compute $\bar{\square}^{**} = \frac{1}{B_2} \cdot \square_{k=1}^{B_2} \hat{\square}^{**k}$
        Compute $\hat{\square}^{**}(\hat{\square}^{*i}) = \sqrt{\frac{1}{B_2-1} \square_{k=1}^{B_2} (\hat{\square}^{**k} - \bar{\square}^{**})^2}$
        Compute $t^*_i = \frac{\hat{\square}^{*i} - \hat{\square}}{\hat{\square}^{**}(\hat{\square}^{*i})}$
next i
Compute $\bar{\square}^{*} = \frac{1}{B_1} \cdot \square_{k=1}^{B_1} \hat{\square}^{*k}$
Compute $\hat{\square}^{*}(\hat{\square}^{*}) = \sqrt{\frac{1}{B_1-1} \square_{k=1}^{B_1} (\hat{\square}^{*k} - \bar{\square}^{*})^2}$
Sort by increasing value the vector $T = \{t^*_1, t^*_2, ..., t^*_{B1}\}$;
Compute the interval of confidence (95%) for $\hat{\square}$ which is:
        $[\ \hat{\square} - t^*_{[.975 \cdot B1]} \cdot \hat{\square}^*(\hat{\square}^*)\ ;\ \hat{\square} - t^*_{[.025 \cdot B1]} \cdot \hat{\square}^*(\hat{\square}^*)\ ]$
```

Figure 2: Algorithm to build confidence interval for a median

Such a scheme does not use any parametric assumption such as Equation 7, and we implicitly build a table for the percentile $t_{\square}$ for constructing the needed confidence interval. The resulting interval is often asymmetrical about 0.

This algorithm has been used to build an approximate confidence interval for the median and the mean as shown in Table 7. For these computations, we set the value of B_1 to 200, and B_2 to 50.

Statistics	CACM s-suffix	CACM Porter	CISI s-suffix	CISI Porter
Median	12.013	14.512	10.919	13.53
$\hat{\square}^*(\hat{\square}^*)$	3.744	3.791	2.592	3.124
$(-t^*_{[.975 \cdot B_1]}, -t^*_{[.025 \cdot B_1]})$	(-2.53, 1.86)	(-2.51, 1.89)	(-2.22, 2.32)	(-2.15, 3.00)
Confidence Interval	[2.55, 18.98]	[5.0, 21.69]	[5.15, 16.94]	[6.81, 22.91]
Mean	19.518	21.003	13.076	13.657
$\hat{\square}^*(\hat{\square}^*)$	3.075	3.006	1.705	1.655
$(-t^*_{[.975 \cdot B_1]}, -t^*_{[.025 \cdot B_1})$	(-1.91, 2.41)	(-1.84, 2.50)	(-1.91, 3.0)	(-1.95, 2.77)
Confidence Interval	[13.64, 26.92]	[15.48, 28.51]	[9.83, 18.20]	[10.43, 18.24]
$S / \sqrt{n}$	3.199	3.200	1.655	1.631
Standard CI	[13.09, 25.94]	[14.57, 27.43]	[9.72, 16.44]	[10.35, 16.97]

Table 7a: Confidence interval for classical Boolean model (□ = 0.05)

Statistics	CACM s-suffix	CACM Porter	CISI s-suffix	CISI Porter
Median	26.994	29.86	17.311	20.877
$\hat{\square}^*(\hat{\square}^*)$	3.881	3.245	3.773	3.679
$(-t^*_{[.975 \cdot B_1]}, -t^*_{[.025 \cdot B_1]})$	(-1.70, 2.45)	(-2.48, 2.18)	(-3.06, 5.11)	(-2.34, 2.35)
Confidence Interval	[20.41, 36.50]	[21.83, 36.95]	[5.78, 36.58]	[12.28, 29.52]
Mean	30.917	32.578	18.597	20.277
$\hat{\square}^*(\hat{\square}^*)$	2.899	2.810	2.276	2.378
$(-t^*_{[.975 \cdot B_1]}, -t^*_{[.025 \cdot B_1]})$	(-1.88, 1.84)	(-1.65, 2.28)	(-1.82, 2.94)	(-2.01, 2.97)
Confidence Interval	[25.48, 36.26]	[27.95, 38.99]	[14.45, 25.28]	[15.50, 27.35]
$S / \sqrt{n}$	2.939	3.036	2.126	2.215
Standard CI	[25.09, 36.74]	[26.48, 38.68]	[14.28, 22.91]	[15.78, 24.77]

Table 7b: Confidence interval for vector-space model (□ = 0.05)

Under the label "Standard CI", one can find the standard interval estimate computed according to Equation 9. The percentile of Student's distribution is $t_{n-1,1-\square} = 2.009$, with $n = 50$ and $\square = 0.05$ ($t_{n-1,1-\square} = 2.03$, with $n = 35$ and $\square = 0.05$). As n -> , the bootstrap and standard intervals converge toward each other, and sample mean examples are provided in Table 7.

The information contained in these tables can be used to answer the question as to how far the guess $\hat{\square}$ might reasonable be. For example, using the CACM collection, the sample mean of the classical Boolean model is 21.003 (Porter's stemming procedure). With a coverage probability of 0.95, we may say that the real mean for this retrieval scheme must be between 15.48 and 28.51 (or between 14.57 and 27.43 using a standard confidence interval).

From these tables, it can be observed that the standard errors associated with the median are greater than those of the sample mean.

3.4 Hypothesis testing

After obtaining a numerical value for the point estimator accuracy and building confidence intervals for them, we then wish to test the null hypothesis H_0 or the validity of the assumption of identical two medians (or means). This assumption will be accepted if two retrieval schemes return statistically similar performances, and rejected if not. Such comparison of treatments or effects represents the major objective of a retrieval experiment.

To achieve this goal, we take a sample of observations $X^p = \{(x_1^a, x_1^b); (x_2^a, x_2^b); ...; (x_k^a, x_k^b); ...; (x_n^a, x_n^b)\}$ representing the average precision at eleven standard recall values for a sample of n queries using respectively the retrieval scheme a (x_k^a) and the search strategy b (x_k^b). We know that each pair of values (x_k^a, x_k^b) ~ i.i.d. of an unknown distribution P. This assertion implies that each x_k^a ~ i.i.d. of F^a, and x_k^b ~ i.i.d. of F^b, where F^a and F^b represent possibly different probability distributions.

If these two retrieval schemes result in similar retrieval effectiveness, we might state the null hypothesis H_0, that the $\text{median}_a = \text{median}_b$ or that the $\text{median}_a - \text{median}_b = 0$ (or $\text{mean}_a - \text{mean}_b = 0$).

To test whether or not the two medians are statistically equal, the first step is to compute a sample $X^d = \{x_1^d, x_2^d, ..., x_k^d, ..., x_n^d\}$ based on the difference between the two search strategies according to the following equation:

$$x_k^d = x_k^a - x_k^b \quad \text{with } x_k^d \sim \text{i.i.d. of } \hat{F}^d.$$

From this sample of size n, we may compute the value $\hat{\square}$, the estimate of our summary statistic □, the median or the mean of the unknown distribution F^d, and we may write the underlying hypotheses of our test as follows:

$$H_0: \square = 0 \quad \text{vs.} \quad H_1: \square \neq 0$$

Based on the numerical value of $\hat{\square}$ which is fixed at its observed value, we compute the sample $U = \{u_1, u_2, ..., u_k, ..., u_n\}$ according to $u_k = x_k^d - \hat{\square}$ so that u_k□~□i.i.d. of $\hat{G}$, and having the null hypothesis distribution, the distribution of $\hat{\square}$ if H_0 is true. On the one hand, if the estimator $\hat{\square}$ represents the median, we might state that $\hat{G}^{-1}(0.5) = 0$, which satisfies H_0. On the other hand, if we have estimated the mean, U represents a sample of a distribution $\hat{G}$ having a mean equal to zero.

Based on the sample U, we may use the general idea of bootstrap methodology. The computer generates a bootstrap sample $U^{*i} = \{u^*_1, u^*_2, ..., u^*_k, ..., u^*_n\}$ of size n drawn from U, with u^*_k ~ i.i.d. of $\hat{G}'$ (with $\hat{G}'^{-1}(0.5) = 0$, or with mean equal to zero).

From each bootstrap sample U^{*i}, we compute a bootstrap replication of the median (or the mean) noted $\hat{\square}^{*i}$. We repeat this process B times and we obtain an empirical distribution of $\hat{\square}^{*i}$ forming the basis for the outcome of our test.

```
Hypothesis testing
Compute $\hat{\square}$, the median (mean) of the empirical distribution
    as $\hat{\square} = t(\hat{F}^{d})$;
Hypothesis: $\square = 0$ (H$_0$) vs. $\square \neq 0$ (H$_1$);
Compute $U = \{u_1, u_2, ..., u_k, ..., u_n\}$ according to $u_k = x_k^{d} - \hat{\square}$;
for i = 1, 2, ..., B (e.g., B = 200) to generate values having the null
  hypothesis distribution;
    Generate $U^{*i} = \{u^*_1, u^*_2, ..., u^*_k, ..., u^*_n\}$ a bootstrap sample;
    Compute $\hat{\square}^{*i}$ the absolute value of the median (mean) of the
        bootstrap sample $U^{*i}$;
next i
Test for equality:
Does the fixed value $|\hat{\square}|$ exceed the threshold value $\hat{\square}^{*i}_{[.95 \cdot B]}$?
    if yes, H$_0$ must be rejected; if not, H$_0$ cannot be rejected.
```

Figure 3: Algorithm determining whether two location statistics are identical

In order to obtain a two-sided test, the absolute values of $\hat{\square}^{*i}$ are sorted. From this empirical distribution, we select a threshold value given a significance level □. This value can be found in the position (1-□) · B and will be compared with the absolute value of the point estimator $\hat{\square}$ of the sample X^d. If the fixed value $|\hat{\square}|$ exceeds the threshold value $\hat{\square}^{*i}_{[1-\square \cdot B]}$, then the null hypothesis H_0 is rejected, leading to the conclusion that the two retrieval schemes present a significant difference. Otherwise, we cannot reject the null hypothesis, inferring that the two search strategies exhibit similar retrieval effectiveness.

As a guideline for choosing an appropriate value for the constant □, we suggest that when □ = 0.01, the resulting test decision can be interpreted as very strong evidence against H_0, while □ = 0.05 represents reasonably strong evidence against H_0 (the value used in Figure 3).

The underlying procedure shown in Figure 3 can also be applied to compute the achieved significance level of the bilateral test (ASL), or the probability of observing the value $\hat{\square}$ when the null hypothesis H_0 is true. This probability is calculated as follows:

$$\text{ASL} = \frac{\square I[|\hat{\square}^{*i}| > |\hat{\square}|]}{B} \qquad \text{with } I[|\hat{\square}^{*i}| > |\hat{\square}|] = \begin{cases} 1 & \text{if } |\hat{\square}^{*i}| > |\hat{\square}| \\ 0 & \text{otherwise} \end{cases}$$

Table 8 demonstrates the bootstrap test results for the equality of two medians, and two means respectively. In rows "Median X^a" and "Median X^b", we have reported the retrieval scheme sample median based on the s-suffix algorithm, and Porter's stemming scheme respectively. The third row indicates the value of the statistic $\hat{\square}$ based on the sample X^d, the difference between the performance obtained with the s-suffix algorithm and Porter's stemming procedure. In the fourth row, we have indicated the absolute value of $\hat{\square}^{*i}_{[.95 \cdot B]}$. Finally, the probability value of ASL is depicted, and the last row denotes the underlying test results. For these computations, the value of B was set to 200.

Statistics	CACM Boolean	CACM vector	CISI Boolean	CISI vector
Median X^a	12.013	26.994	10.919	17.311
Median X^b	14.512	29.863	13.53	20.877
$\hat{\square}$	0.0	-0.697	0.034	-0.199
$\hat{\square}^{*i}_{[.95 \cdot B]}$	0.0	4.376	0.282	1.391
ASL	1.0	0.495	0.45	0.6
Decision	H_0 cannot be rejected: 0.05	H_0 cannot be rejected: 0.05	H_0 cannot be rejected: 0.05	H_0 cannot be rejected: 0.05
Mean X^a	19.518	30.917	13.076	18.597
Mean X^b	21.003	32.578	13.657	20.277
$\hat{\square}$	-1.485	-1.661	-0.581	-1.679
$\hat{\square}^{*i}_{[.95 \cdot B]}$	2.123	3.552	1.298	1.386
ASL	0.18	0.37	0.415	0.02
Decision	H_0 cannot be rejected: 0.05	H_0 cannot be rejected: 0.05	H_0 cannot be rejected: 0.05	H_0 rejected $\square = 0.05$

Table 8: Testing the equality of two medians or two means when comparing stemming procedures (two-sided test, $\square = 0.05$)

The resulting decision that can be inferred form this table is that the s-suffix stemming scheme performs in a manner similar to Porter's stemming procedure.

However, our data can also be used to verify the null hypothesis H_0 that the classical Boolean model performs similarly to the vector-space model based on the cosine coefficient. Of course, such an assumption can be viewed as a devil's advocate because we already know that the vector-processing retrieval strategy performs better than the Boolean model. To investigate this hypothesis, we calculate the difference between the Boolean model average precision against that achieved with the vector-space model. Of course, these comparisons are based on the same suffix-stripping algorithm (see Table□9).

Statistics	CACM s-suffix	CACM Porter	CISI s-suffix	CISI Porter
Median X^a	12.013	14.512	10.919	13.53
Median X^b	26.994	29.863	17.311	20.877
$\hat{\square}$	-9.221	-9.268	-4.265	-5.952
$\hat{\square}^{*i}_{[.95 \cdot B]}$	3.930	6.308	3.801	4.3
ASL	0.0	0.0	0.01	0.005
Decision	H_0 rejected $\square = 0.05$	H_0 rejected $\square = 0.05$	H_0 rejected $\square = 0.05$	H_0 rejected $\square = 0.05$
Mean X^a	19.518	21.003	13.076	13.657
Mean X^b	30.917	32.578	18.597	20.277
$\hat{\square}$	-11.399	-11.575	-5.521	-6.620
$\hat{\square}^{*i}_{[.95 \cdot B]}$	5.649	6.830	3.218	3.222
ASL	0.0	0.0	0.0	0.0
Decision	H_0 rejected $\square = 0.05$	H_0 rejected $\square = 0.05$	H_0 rejected $\square = 0.05$	H_0 rejected $\square = 0.05$

Table 9: Testing the equality of two medians or two means when comparing two retrieval models (two-sided test, $\square = 0.05$)

The conclusion that can be drawn from Table□9 is clear: the vector-space model performs in a manner superior to the classical Boolean model using either the mean or the median as a location statistic, or the suffix-stripping algorithm.

4. CONCLUSION

The aim of this paper was to evaluate the retrieval effectiveness of a search system and to form a firm theoretical basis for comparing retrieval schemes. After reviewing traditional statistical tests used in information retrieval studies, we suggest rejecting the paired t-test to verify whether or not a retrieval scheme is better than another, because the underlying distribution of the data does not always follow a normal distribution. Moreover, since the hypotheses underlying nonparametric tests are not always strictly respected in the information retrieval domain (van Rijsbergen, 1979), this study suggests using the bootstrap methodology to both analyze the performance of a single retrieval mechanism and to compare two search strategies. However, the bootstrap approach is not an "assumption-free" method and requires that the observations are independent and identically distributed (i.i.d.). In information retrieval, this means that we must assume that the queries sample associated with a given test collection is a reasonable representative of the requests population.

The bootstrap methodology retains the advantage of relieving the investigator from having to make assumptions imposed by both parametric and nonparametric statistical models, or having to derive formulae that can be hard to come by. This paper explains how the bootstrap resampling approach can be applied to building a confidence interval for a given statistic (e.g., the mean or the median) and developing a technique for the application of this approach for statistical inferences. This paper also suggests using the sample median instead of the sample mean as a location measure for information retrieval data.

Even if this study proposes the use of average precision at eleven recall values as a measure of the retrieval performance of a search strategy, the underlying methodology can be applied to other measures of retrieval effectiveness such as the fallout ratio, the expected search length (Cooper, 1968), etc.

Our evaluation methodology indicates whether or not a difference between two retrieval techniques can be considered as significant. However, real retrieval systems are ultimately judged by users, and for them, even a difference that cannot be considered as significant by a statistical test may be both valuable and important if it occurs repeatedly in various contexts (Keen, 1992). Moreover, tests based on small test collections might not always reflect retrieval performance in very large commercial full-text environments (Ledwith, 1992). However, this criticisms and other related to test collections, do not invalidate most of the important conclusions that can be drawn from retrieval experiments (Salton, 1992).

Acknowledgments: I would like to thank Prof. C. Léger from the University of Montréal for his useful help in the comprehension of the aims and principles of the bootstrap methodology. I would also thank V. Rousson for his reading and commenting of an earlier version of this paper. This research was supported by the SNSF (Swiss National Science Fondation) under grants 21-37'345.93 and 20-43'217.95.

REFERENCES

Antille, A., Kersting, G., & Zucchini W. (1982). Testing symmetry. *Journal of the American Statistical Association,* 77(379), 639-646.

Bickel, P. J., & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6), 1196-1217.

Boyce B. R., Meadow C. T., & Kraft D. H. (1994). *Measurement in information science.* San Diego, CA: Academic Press.

Cleverdon, C. W. (1984). Optimizing convenient on-line access to bibliographic databases. *Information Service & Use*, 4, 37-47.

Conover, W. J. (1980). *Practical nonparametric statistics*. 2nd ed., New-York, NY: John Wiley & Sons.

Cooper, W. S. (1968). A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19(1), 30-41.

Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science,* 24, 87-100.

Efron, B., & Tibshirani, R. J. (1986). Bootstrap measures for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1, 54-77.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New-York, NY: Chapman & Hall.

Fienberg, S. E. (1989). *The analysis of cross-classified categorical data*. Cambridge, MA: The MIT Press.

Fox, E. A. (1983). *Characterization of two experimental collections in computer and information science containing textual and bibliographic concepts*. Technical Report TR 83-561, Department of Computer Science, Cornell University.

Freeman, D. E. (1987). *Applied categorical data analysis*. New-York, NY: Marcel Dekker.

Frakes, W. B. (1992). Stemming algorithms. In W. B. Frakes, R. Baeza-Yates (Ed.), *Information retrieval, data structures & algorithms* (pp. 131-160), Englewood Cliffs, NJ: Prentice-Hall.

Gluck, M. (1996). Exploring the relationship between user satisfaction and relevance in information systems. *Information Processing & Management*. 32(1), 89-104.

Grimm, L. G. (1993). *Statistical applications for the behavioral sciences*. New-York, NY: John Wiley & Sons.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society of Information Science*, 42(1), 7-15.

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science* 47(1), 37-49.

Herring, C., Palmore, J. I. (1995). Random number generators are chaotic. *Communications of the ACM*, Technical Correspondence, 38(1), 121-122.

Hull, D. (1993, June). *Using statistical testing in the evaluation of retrieval experiments.* Proceedings of the 16th International Conference of the ACM-SIGIR'93, Pittsburgh, PA, 329-338.

Hull, D. (1996). Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*. 47(1), 70-84.

Keen, M. E. (1992). Presenting results of experimental retrieval comparisons. *Information Processing & Management*, 28(4), 491-502.

Kraft D. H., & Boyce, B. R. (1991). *Operations research for libraries and information agencies.* San Diego, CA: Academic Press.

Krovetz, R. (1993, June). *Viewing morphology as an inference process*. Proceedings of the 16th International Conference of the ACM-SIGIR'93, Pittsburgh, PA, 191-202.

L'Écuyer, P., & Côté, S. (1991). Implementing a random number package with splitting facilities. *ACM Transactions on Mathematical Software*, 17(1), 98-111.

Ledwith, R. (1992). On the difficulties of applying the results of information retrieval research to aid in the searching of large scientific databases. *Information Processing & Management*, 28(4), 451-455.

Léger, C., Politis, D. N., & Romano, J. P. (1992). Bootstrap technology and applications. *Technometrics*, 34(4), 378-398.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1), 22-31.

Paice, C. D. (1994, July). *An evaluation method for stemming procedure*. Proceedings of the 17th International Conference of the ACM-SIGIR'94, Dublin, Ireland, 42-50.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.

van Rijsbergen, C. J. (1979). *Information retrieval*. 2nd ed., London, UK: Butterworths.

Saracevic, T. (1975). Relevance: A review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, 26, 321-343.

Saracevic, T., Kantor, P., Chamis A. Y. & Trivison D. (1988). A study of information seeking and retrieving. I Background and methodology. *Journal of the American Society for Information Science*, 39, 161-176.

Saracevic, T. (1995, July). *Evaluation of evaluation in information retrieval.* Proceedings of the 18th International Conference of the ACM-SIGIR'95, Seattle, WA, 137-146.

Salton, G. & McGill, M. J. (1983). *Introduction to modern information retrieval.* New-York, NY: McGraw-Hill.

Salton, G. (1989). *Automatic text processing, the transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

Salton, G. (1992). The state of retrieval system evaluation. *Information Processing & Management*, 28(4), 441-449.

Savoy, J. (1993). Stemming of french words based on grammatical category. *Journal of the American Society for Information Science*, 44(1), 1-9.

Savoy, J. (1994). A learning scheme for information retrieval in hypertext. *Information Processing & Management*, 30(4), 515-533.

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*, 29, 3-48.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*, New-York, NY: McGraw-Hill.

Sparck Jones, K., & Bates, R. G. (1977). *Research on automatic indexing 1974-1976*. Technical Report, Computer Laboratory, University of Cambridge (UK).

Tague-Sutcliffe, J., & Blustein, J. (1994, November). *A statistical analysis of the TREC-3 data*. Proceedings of the 3rd Text REtrieval Conference TREC'3, Gaithersburg, MD, 385-398.

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revised. *Information Processing & Management*, 28(4), 467-490.

Tezuka, S., & L'Écuyer, P. (1991). Efficient and portable combined Tausworthe random number generators. *ACM Transactions on Modeling and Computer Simulation*, 1(2), 99-112.